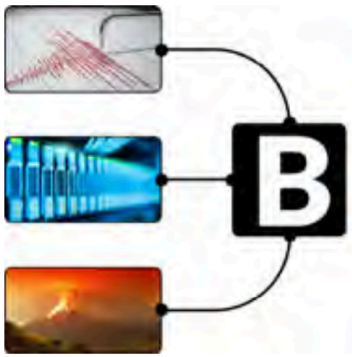


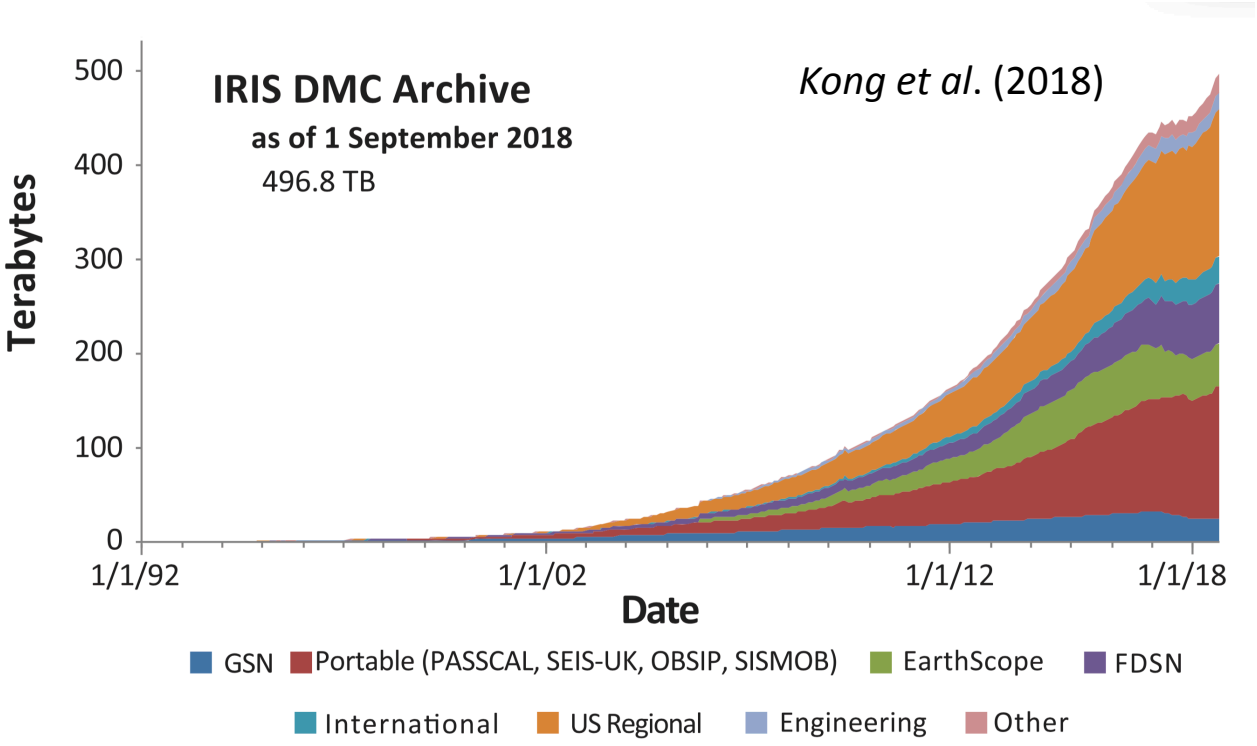
BEYOND THE BLACK BOX

The Future of Machine Learning and Data-Intensive Computing in the Solid Earth Geosciences

Summary of Current Status and Opportunities for Machine Learning and Data Intensive Computing in the Geosciences

Greg Beroza, Stanford University

Need to work with massive data volumes



A single DAS experiment can generate this much data in a few months.

How are we going to process this data?

Can we move this much data around? Should we even try?

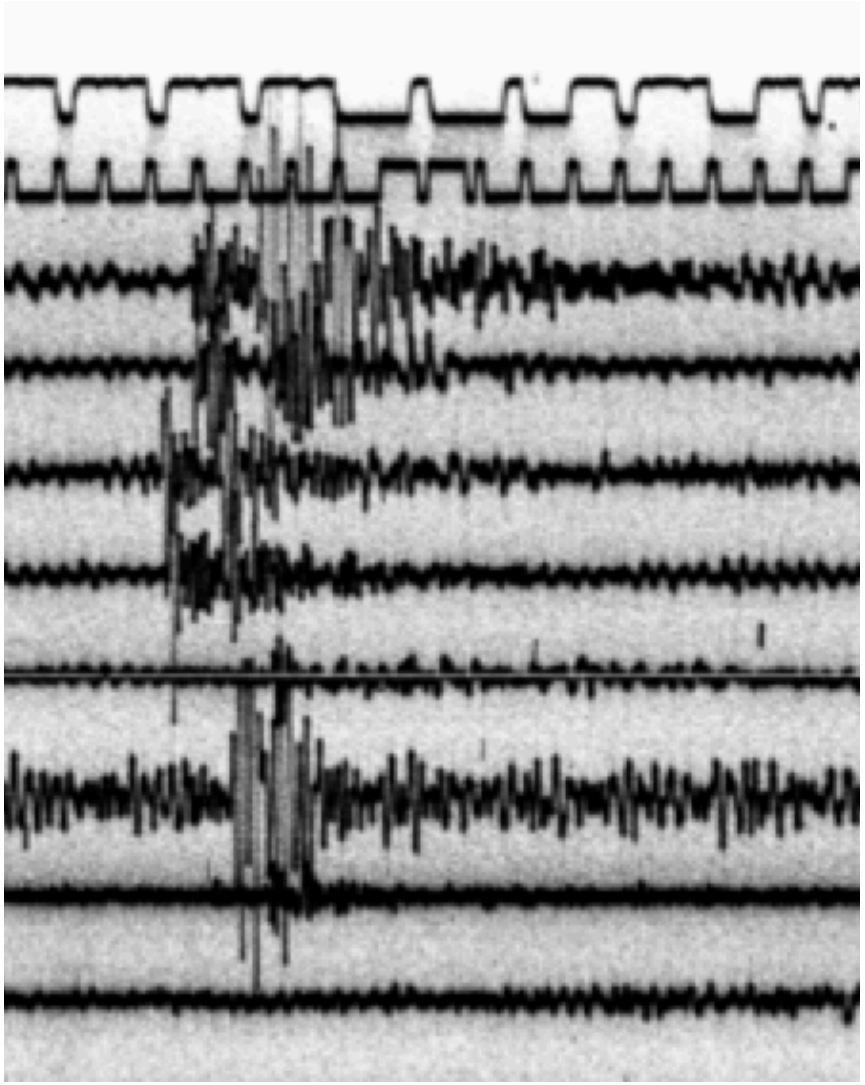
Dark (Legacy) Data

~100 years

Analog data from 1889

~33 years

Digital data standard by the 80s



Time increases left to right (time code is at top) and each line is a channel

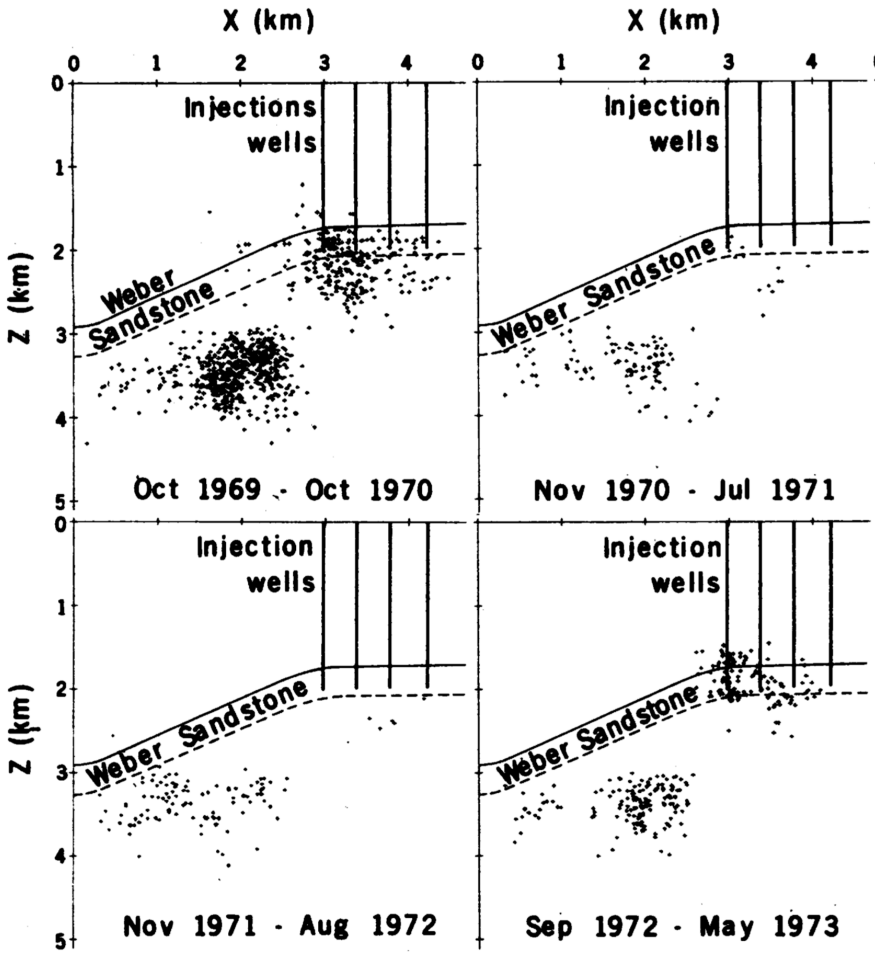
Lines overlap when things get interesting – the bigger the earthquake the greater the overlap (and the fainter the trace).

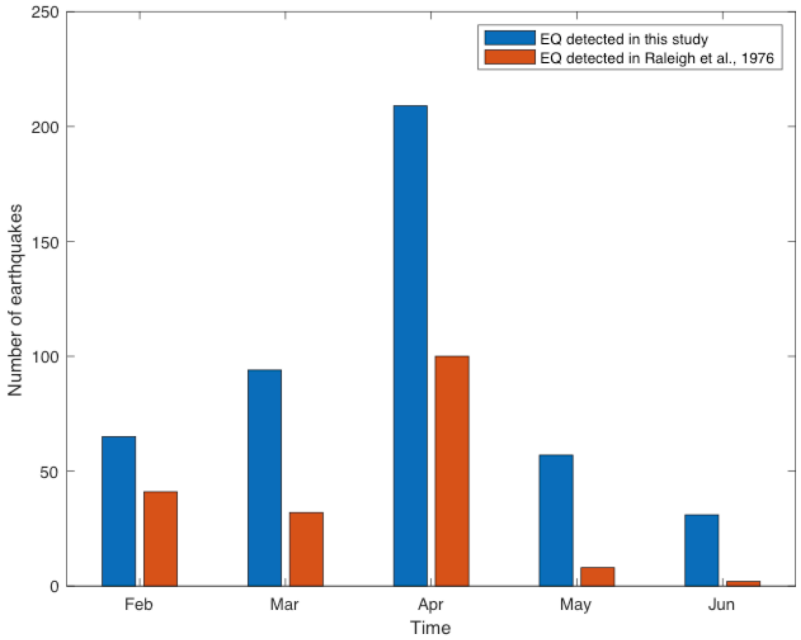
Difficult to disentangle traces to get time series that we can analyze by standard methods.

Work with the Image
(avoid vectorization)

Wang et al. (2018)

Revisiting the Rangely Earthquake Control Experiment

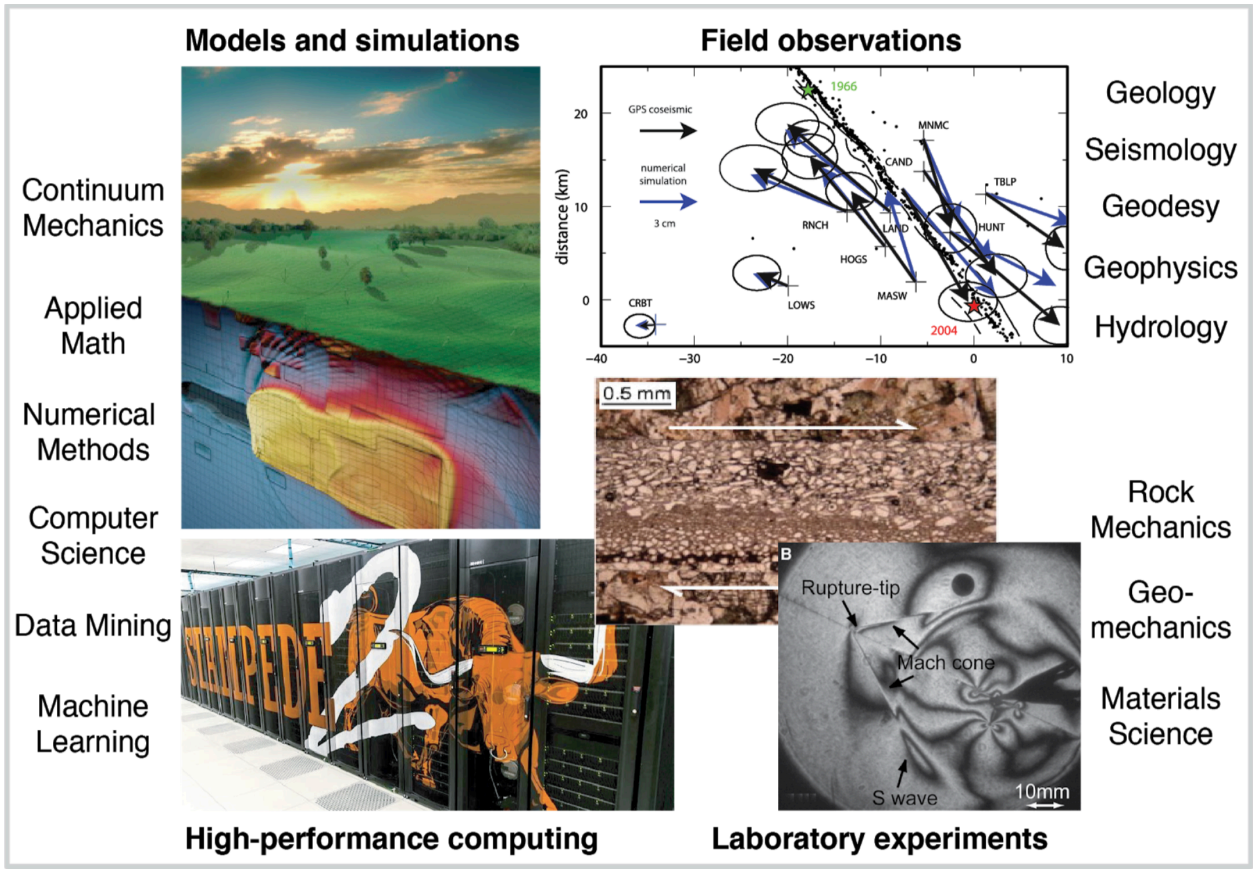




Number of earthquakes detected by the neural network in this study(Feb. to Jun. 1973) compared with the number of events reported in Raleigh et al., 1976.

Wang et al. (2018)

How to diffuse AI through the geosciences?



Lapusta et al. (2019)

Seismology → Earthquake Science

- Network seismology
- LiDAR topography
- GNSS and InSAR
- Simulation/Modeling
- Exploring relationships (e.g., forecasting)

How to choose the right approach?

PageRank for Earthquakes: cluster similar waveforms to extract LFEs from tremor (*Aguiar and Beroza,* 2014)

FAST: Data mining for repeating signals without templates (*Yoon et al.,* 2015)

Unsupervised: exploits similarity in unlabeled data

CRED: Deep learning for earthquake detection (*Mousavi et al.,* 2019)

DeepDenoiser: Deep learning for denoising (*Zhu et al.,* 2019)

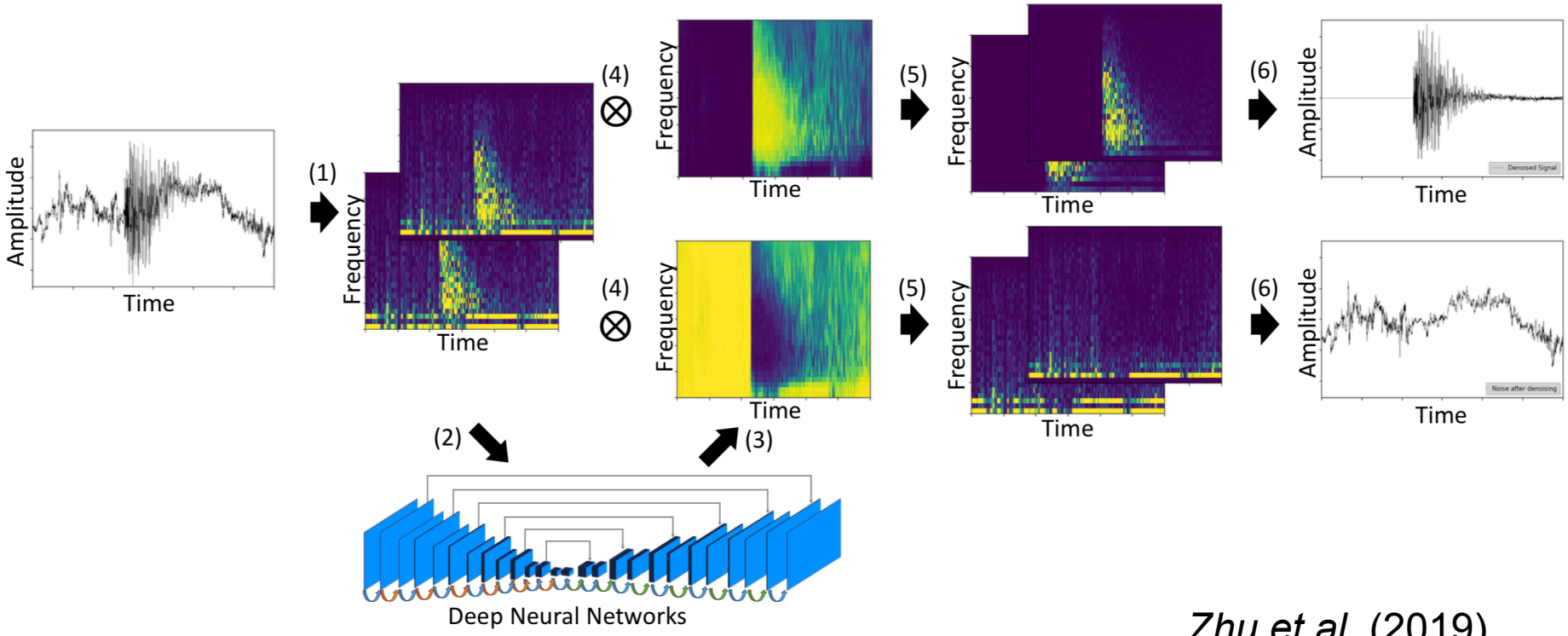
PhaseNet: Machine learning for phase picking (*Zhu et al.,* 2019)

Supervised: learns from labeled data

Deep Autoencoder: Deep learning to discriminate earthquake types with few data. (*Mousavi et al.,* 2019)

Self-supervised: reduces dimensionality

DeepDenoiser learns signal and noise



Zhu et al. (2019)

Signal – Noise Separation

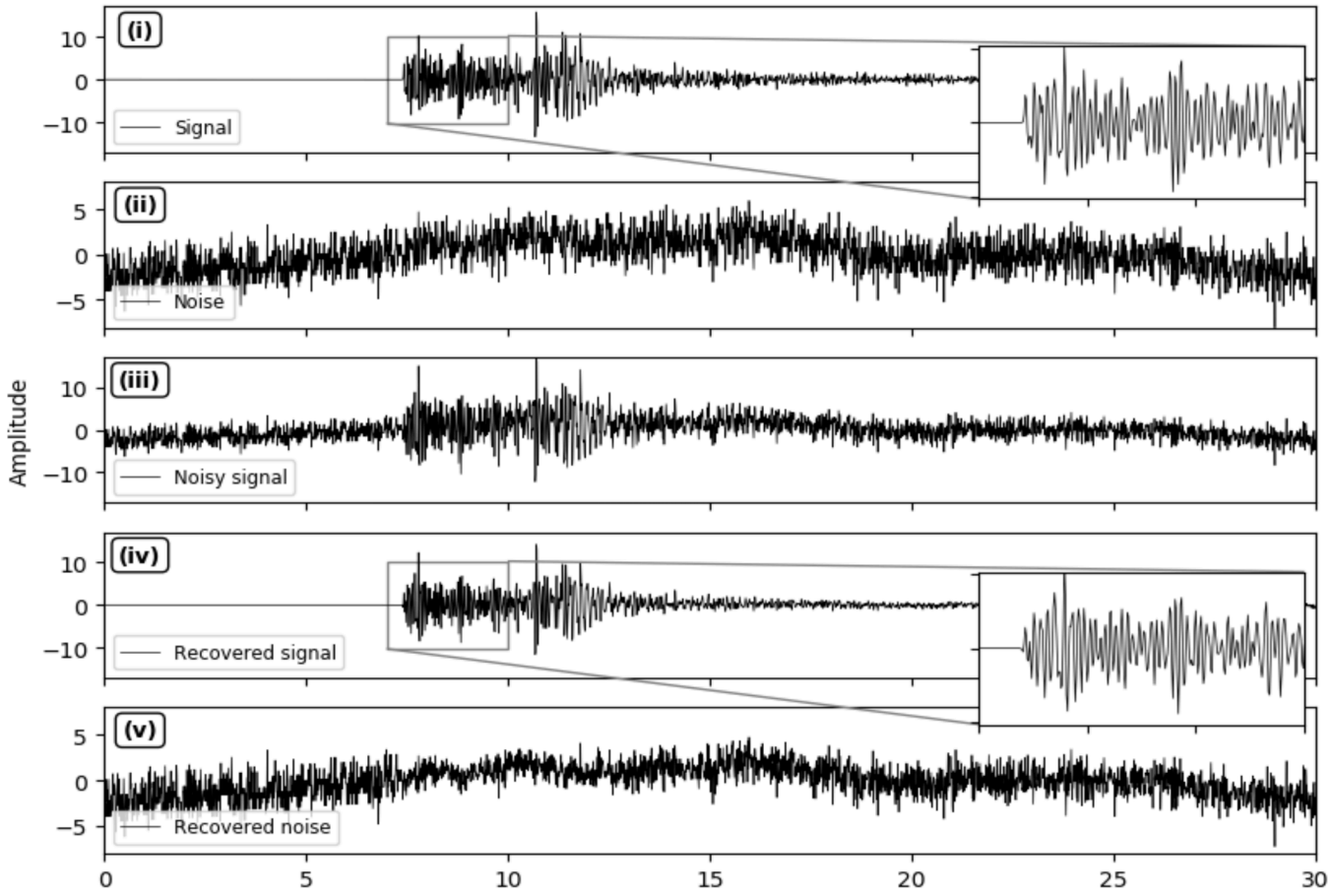
Clean Signal

Noise

Sum

Recovered Signal

Recovered Noise



Zhu et al. (2019)

Some potential applications

Conventional Seismic Monitoring
Other Analysis (e.g. receiver functions)
Urban Seismic Monitoring
- MeSONet (Tokyo)
- Nodal Array data (Long Beach)
- DAS

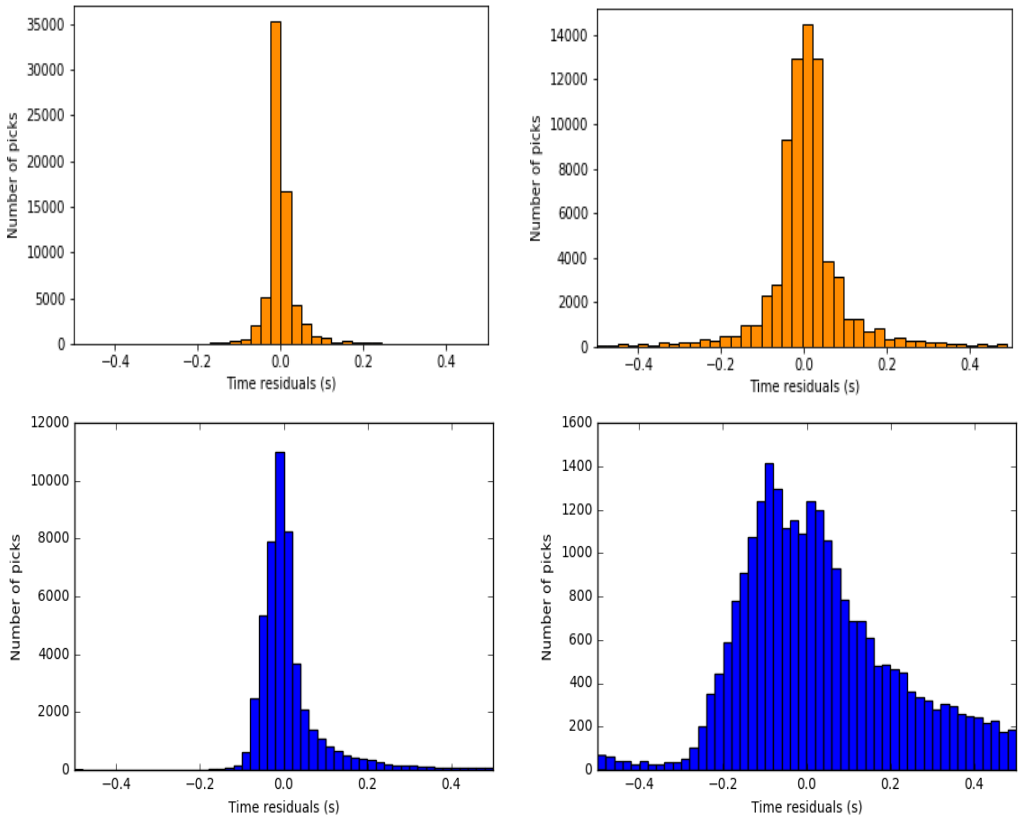
Seafloor Seismic Monitoring
- OBS
- S-Net (Japan)
- DAS

Volcano Monitoring
Also useful for “de-signaling”

P S

PhaseNet

Standard Approach



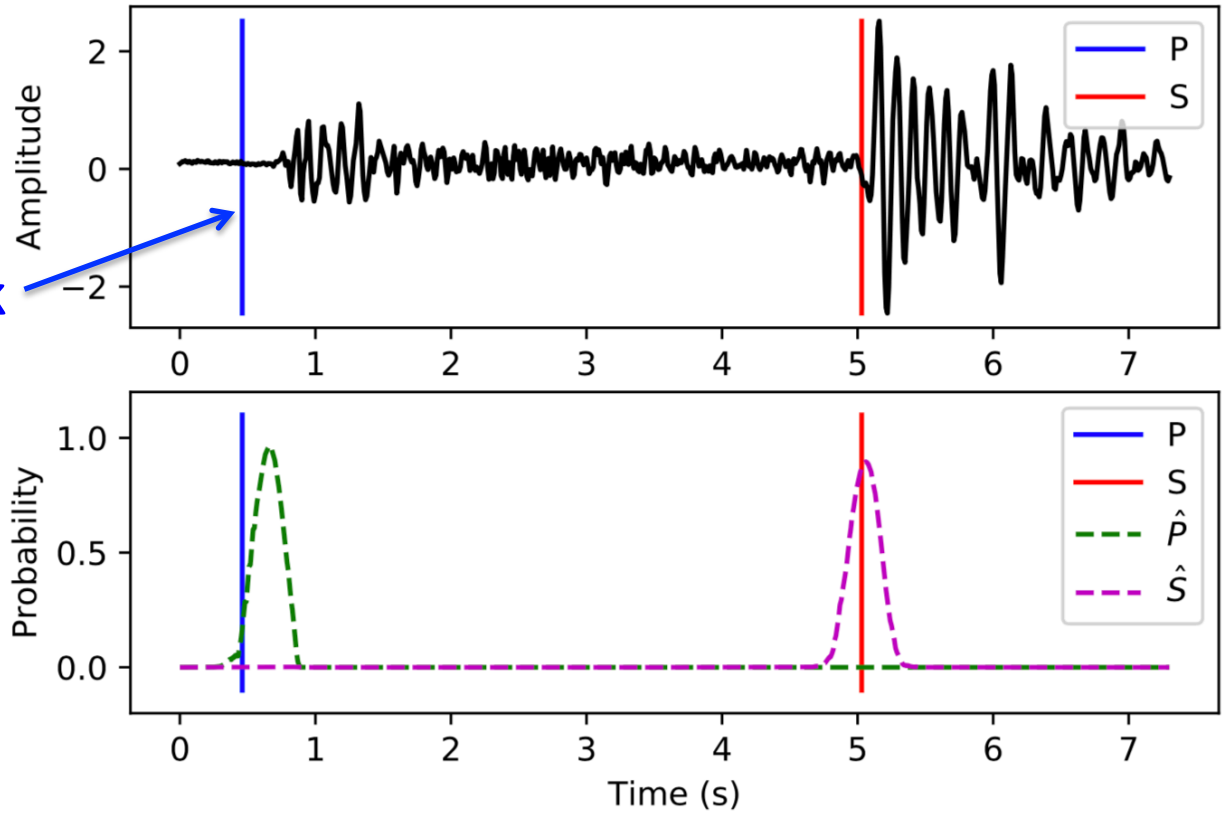
Zhu and Beroza (2019)

Compared with reviewed picks – taken as correct

What is ground truth?

Ground truth? Analyst-reviewed picks have errors

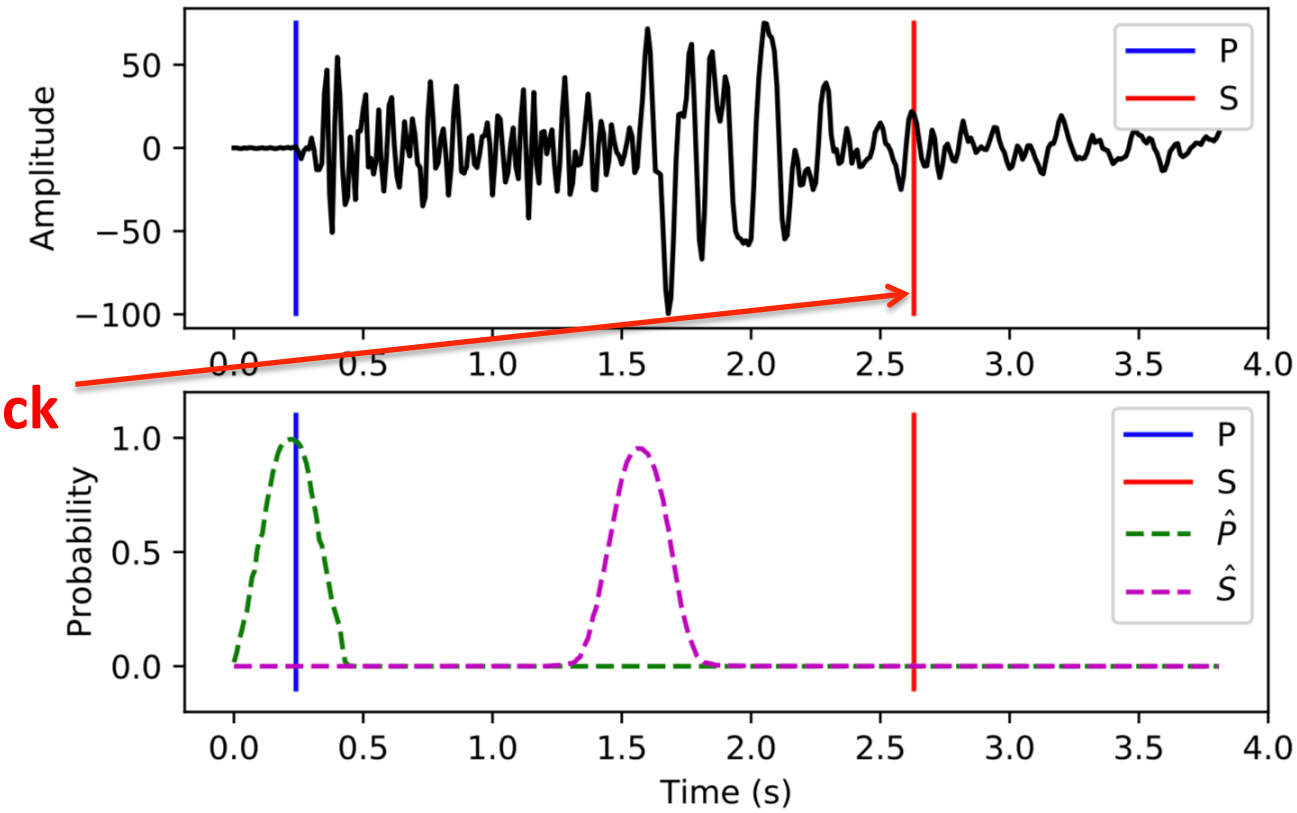
Example of bad P pick



Zhu and Beroza (2019)

Ground truth? Analyst-reviewed picks have errors

Example of bad S pick



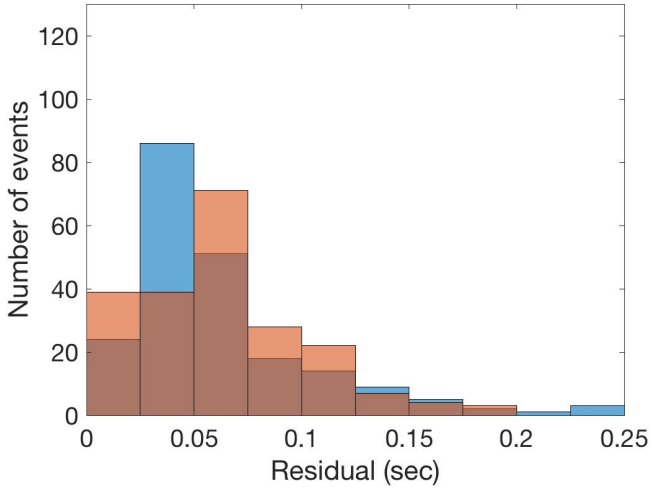
Zhu and Beroza (2019)

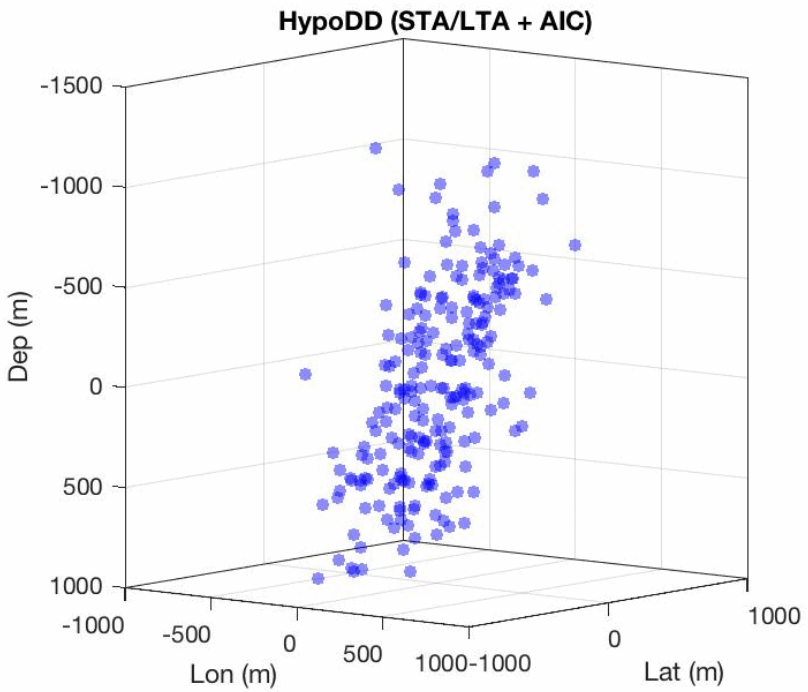
Test of PhaseNet on data from Apennines

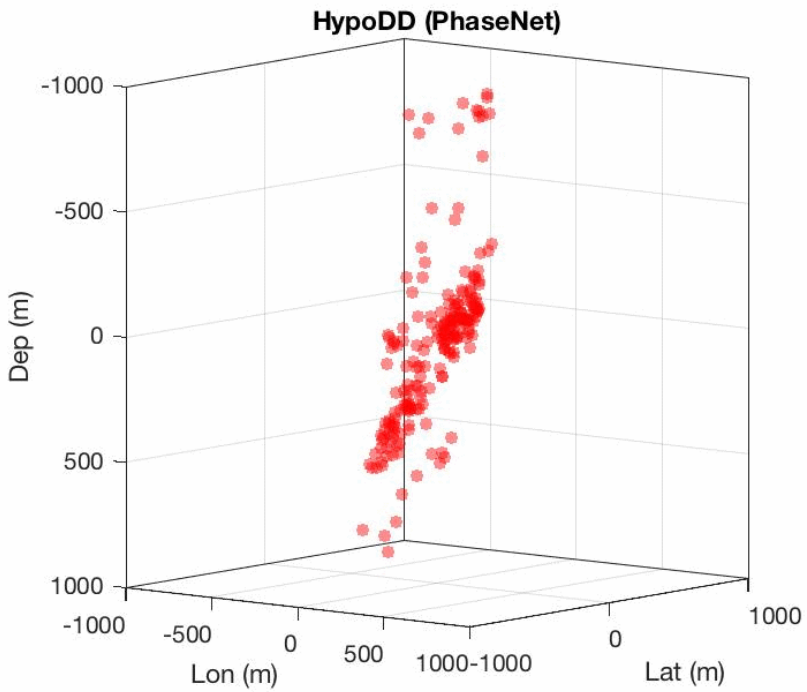
Test cluster

PhaseNet: 52,882 picks
STA/LTA + AIC: 26,306 picks

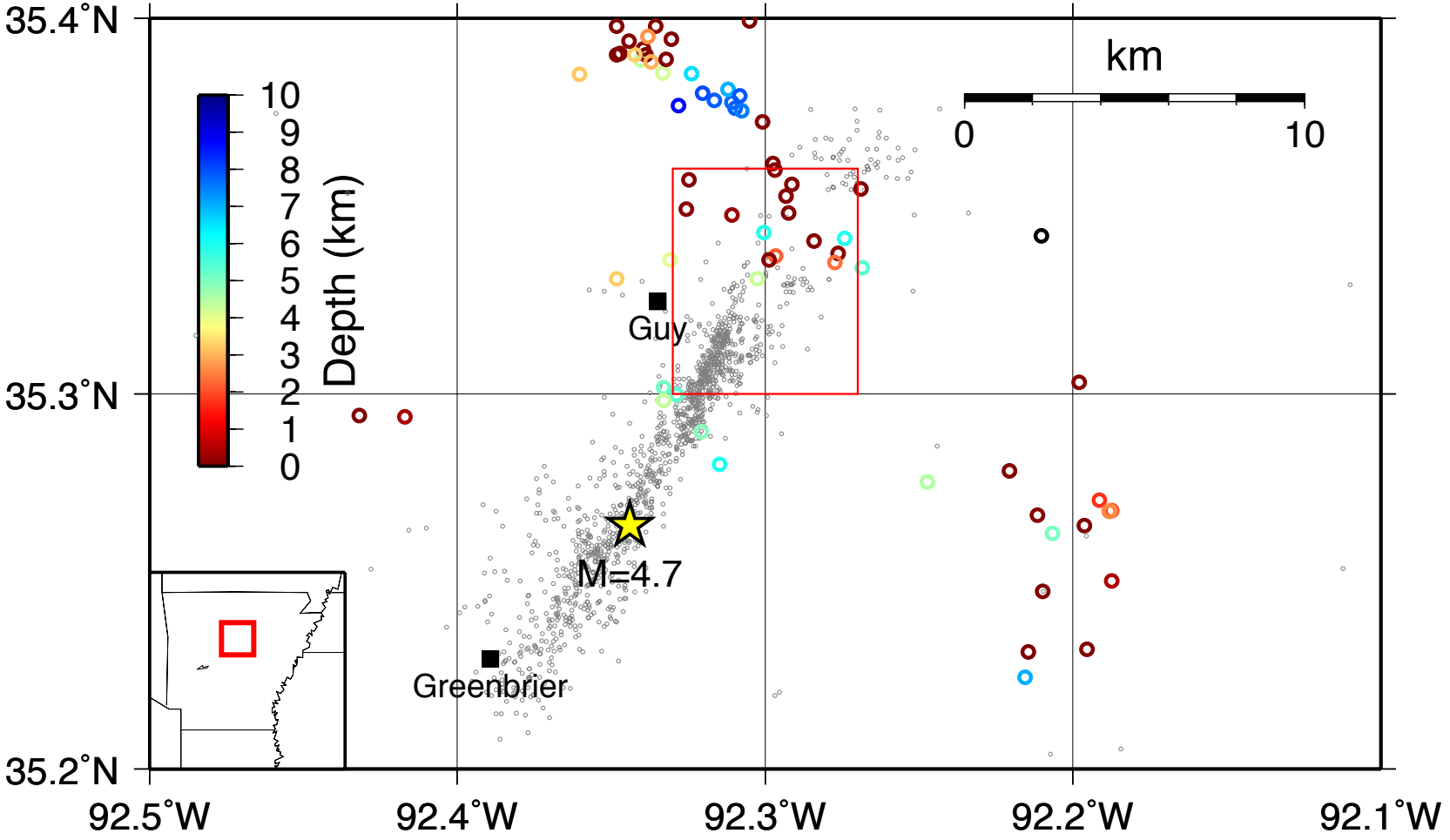
Larger residuals, but more picks
(S waves) = better locations



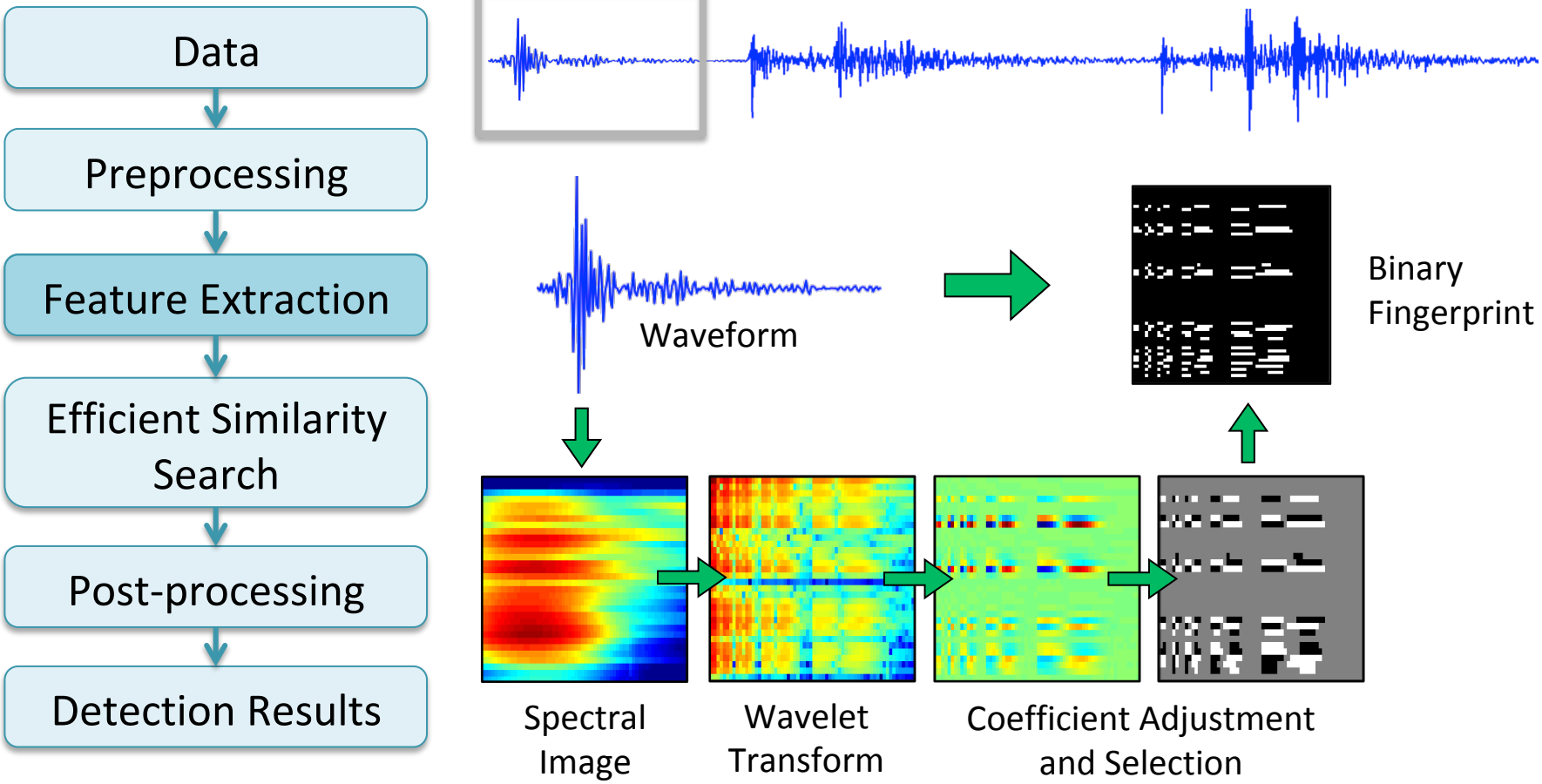

HypoDD (STA/LTA + AIC)
-1500
-1000
-500
0
500
1000
Dep (m)
-1000
-500
0
500
1000
Lon (m)
-1000
0
1000
Lat (m)


HypoDD (PhaseNet)
-1000
-500
0
500
1000
Dep (m)
-1000
-500
0
500
1000
Lon (m)
-1000
0
1000
Lat (m)

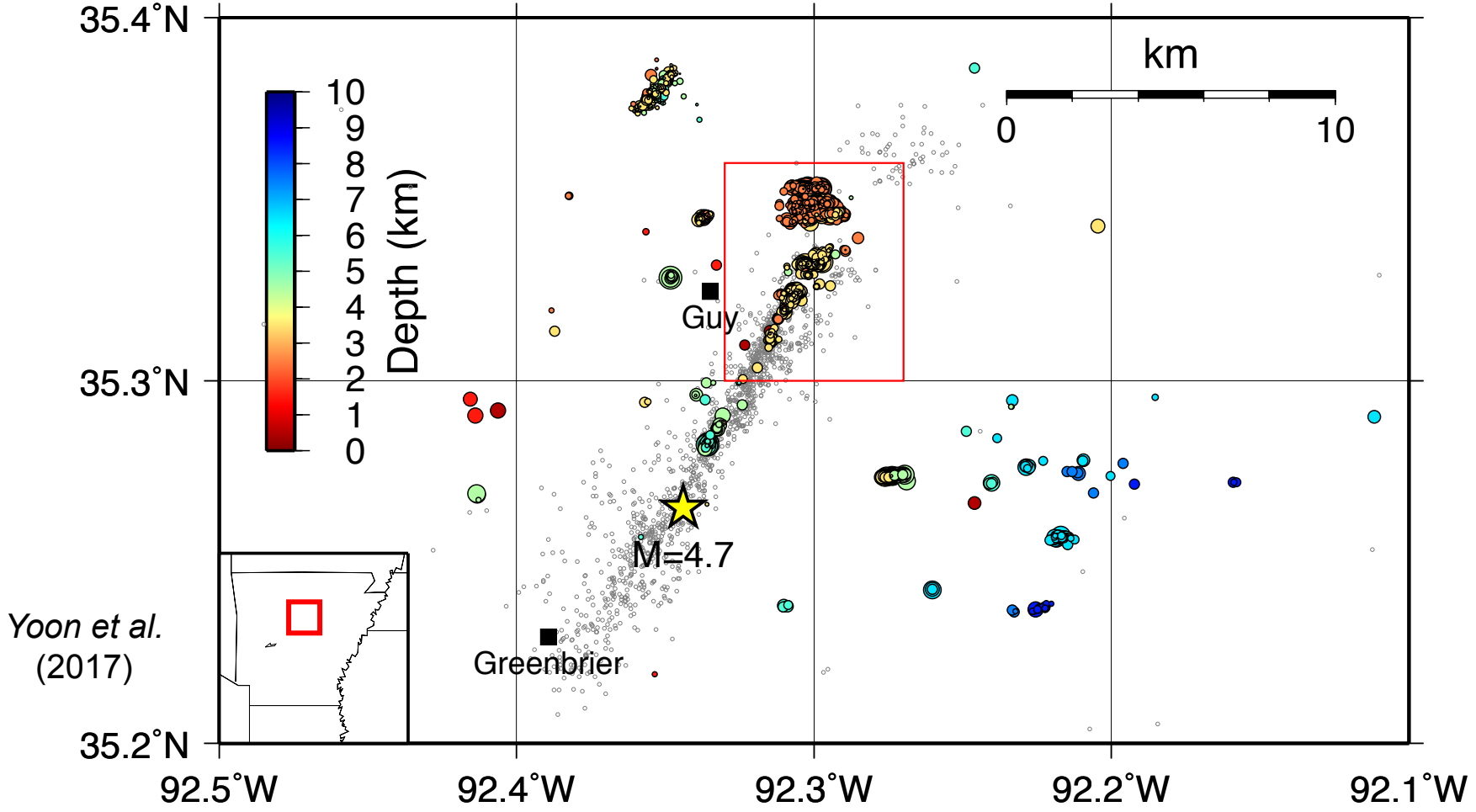
Before: 75 catalog earthquakes, 1.2 < M_L < 2.9



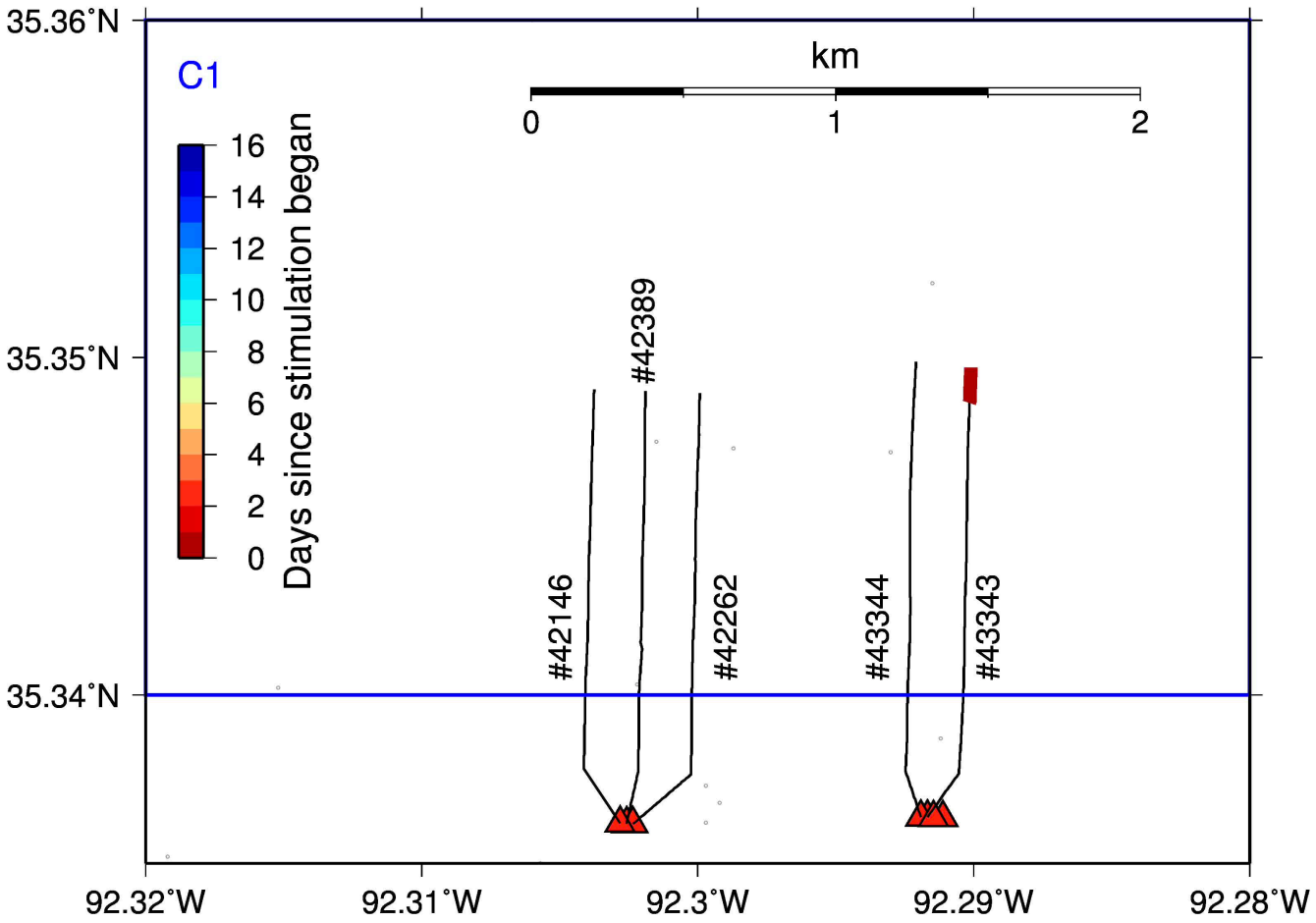
FAST Detection Pipeline



After: 14,604 detected earthquakes, -1.5 < M_L < 2.9



Yoon et al. (2017)



Yoon et al. (2017)

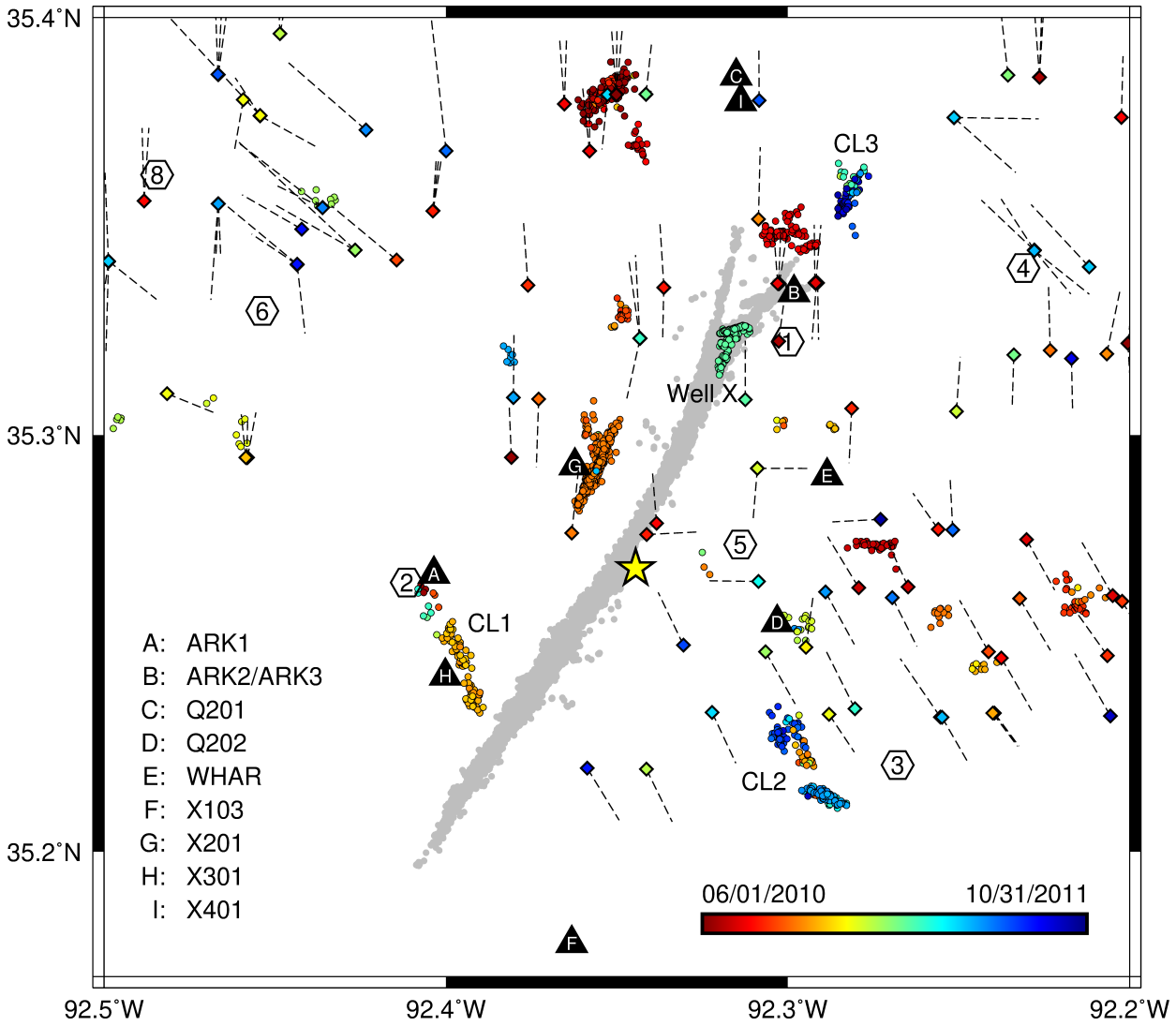
Generalization?

Similarity Search for Earthquakes

Informed Search: Template matching or subspace projection of known event waveforms.

Uninformed Search: Discovery of templates through naïve correlation, Pagerank clustering, or approximate search by LSH.

Generalized Similarity Search: Generalization of strict similarity search to more permissive similarity in waveform characteristics using machine learning.



ML−based catalog

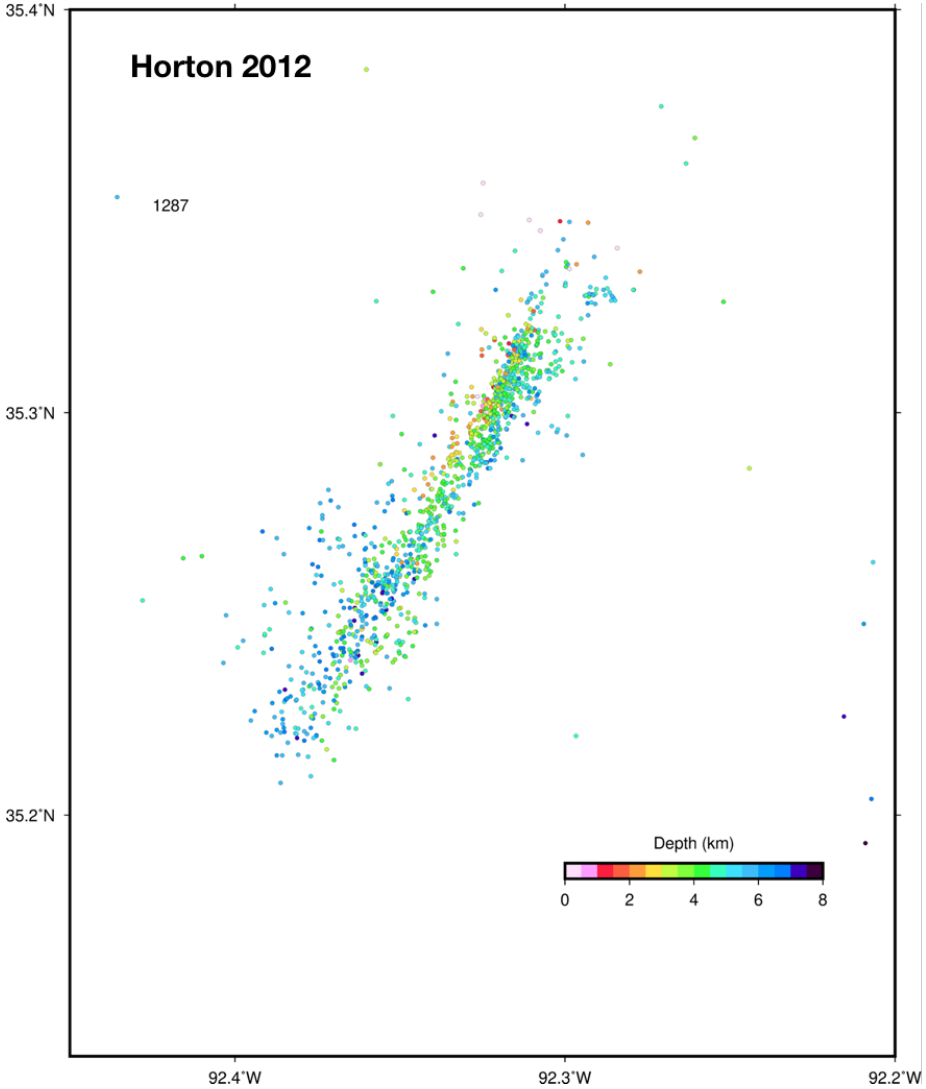
PhaseNet trained on NCSN data only − generalizes well

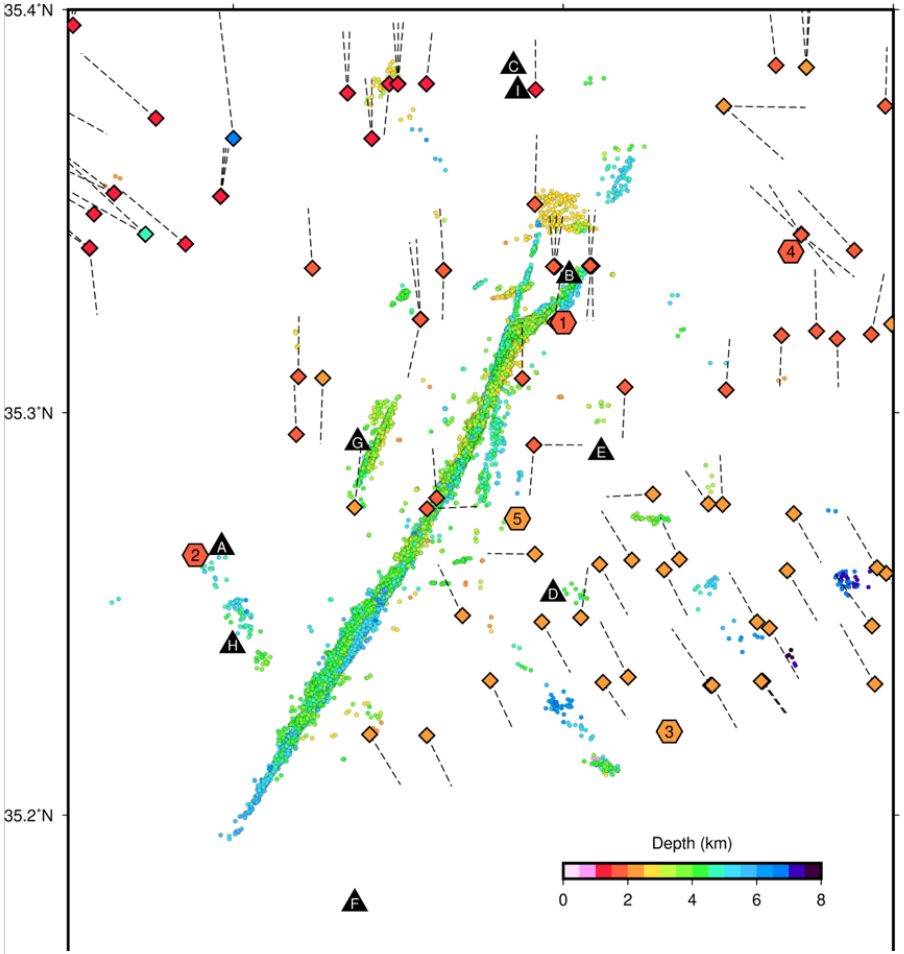
~90,000 events

Event triggering

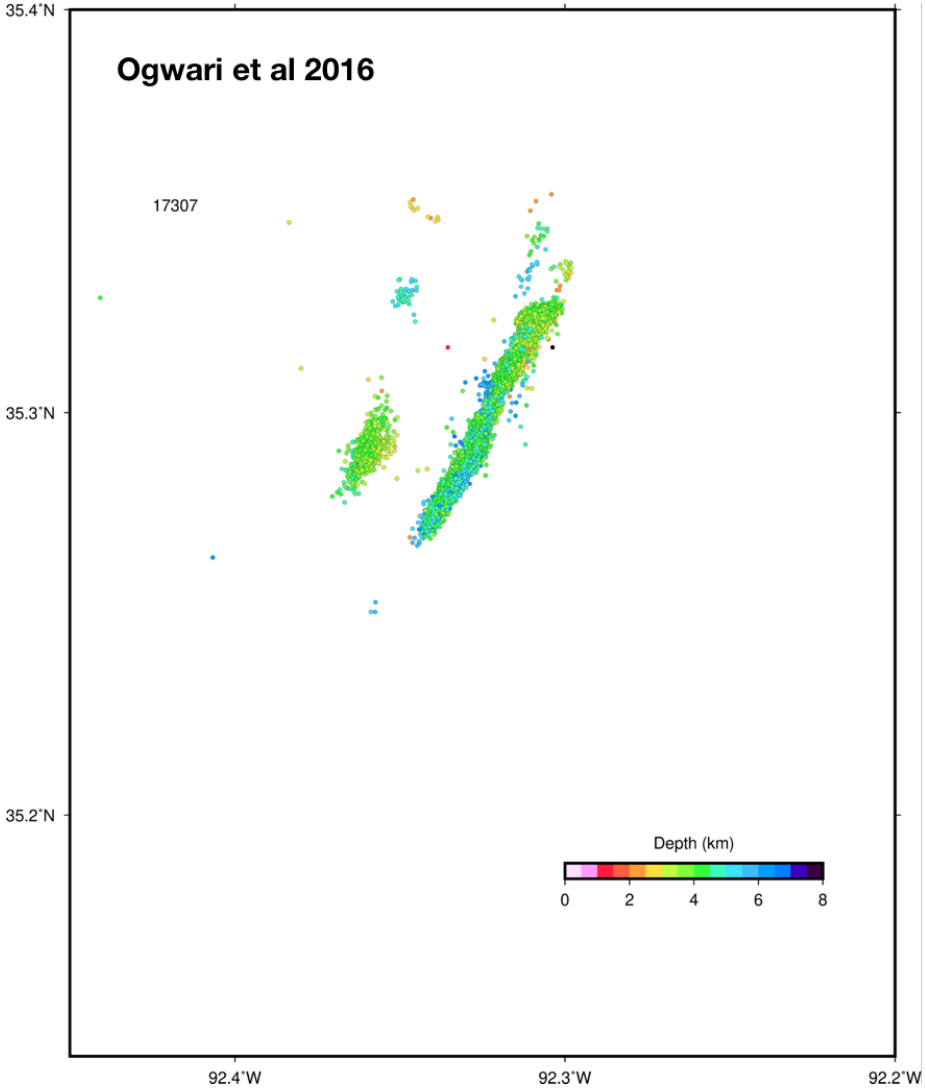
- hydraulic stimulation
- deep disposal wells

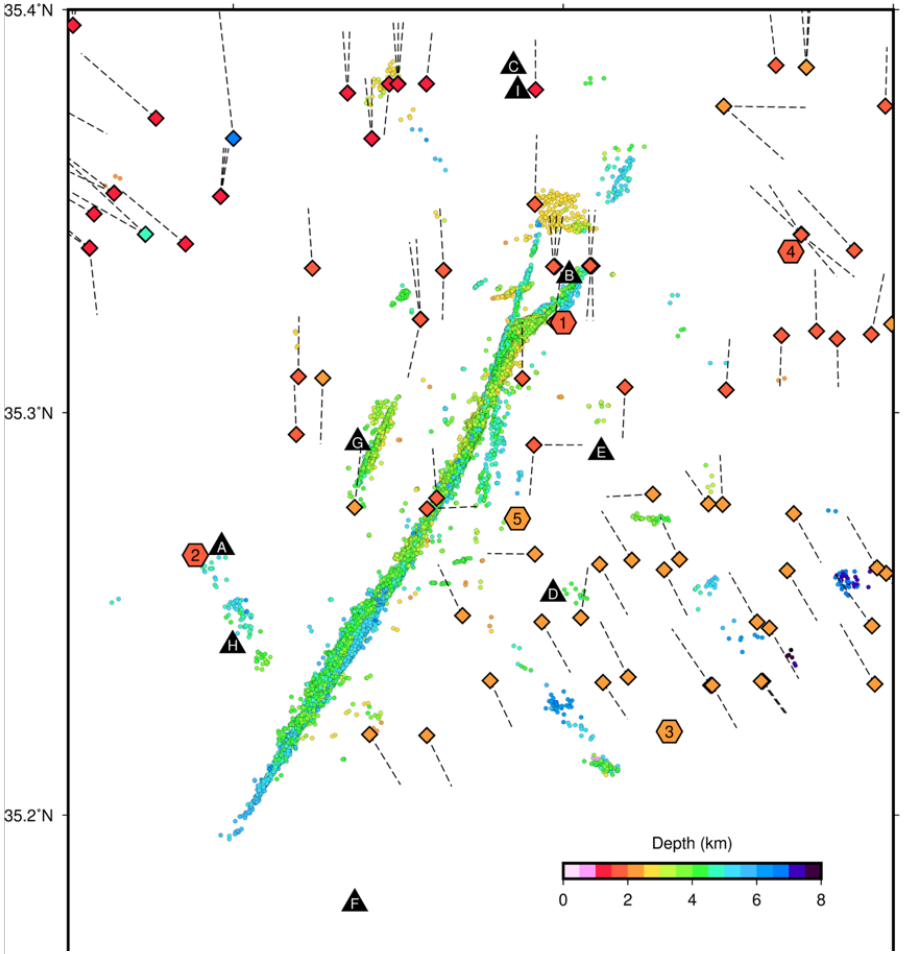
Park et al. (2019)



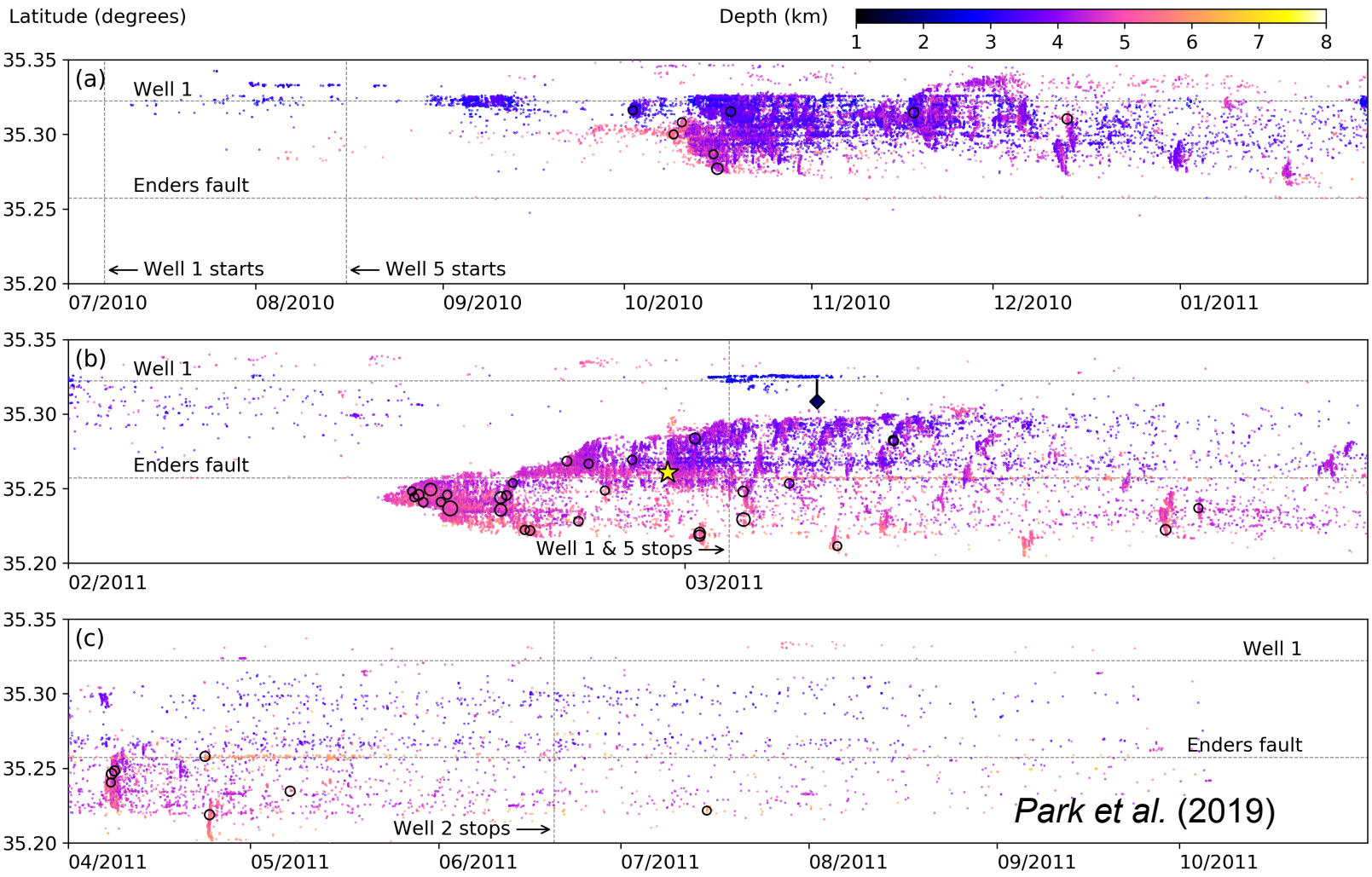


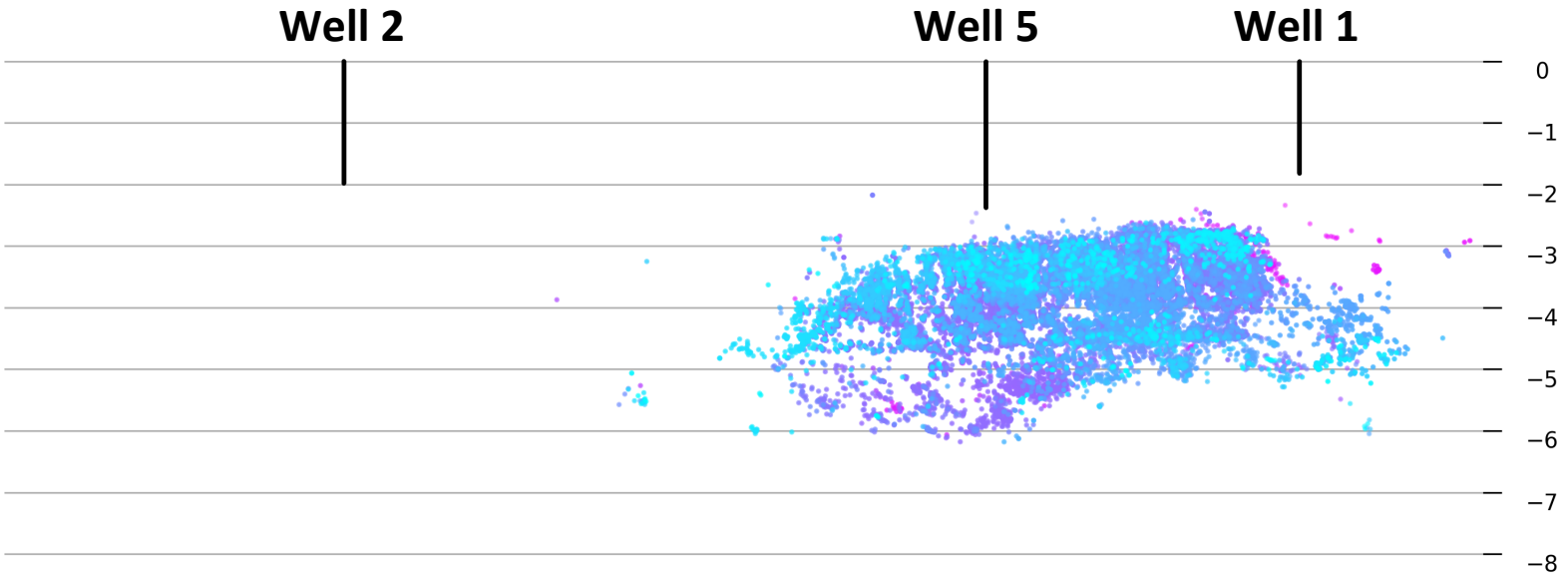
Park et al. (2019)



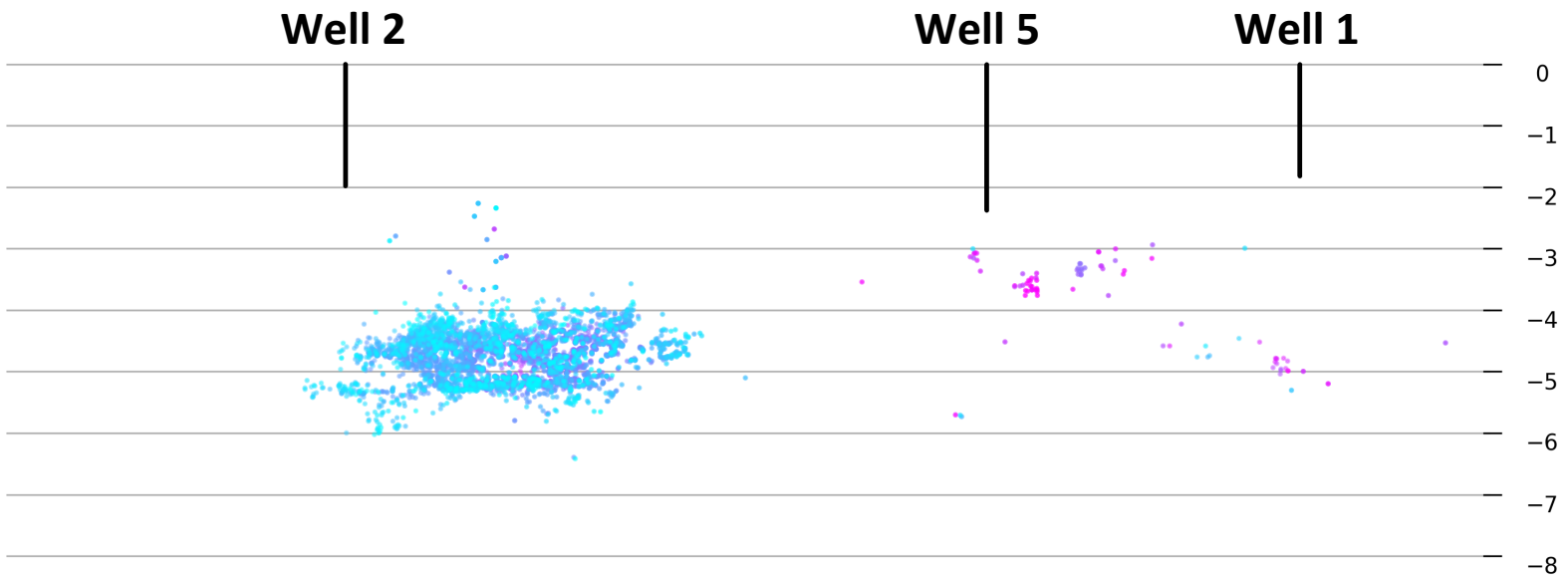


Park et al. (2019)

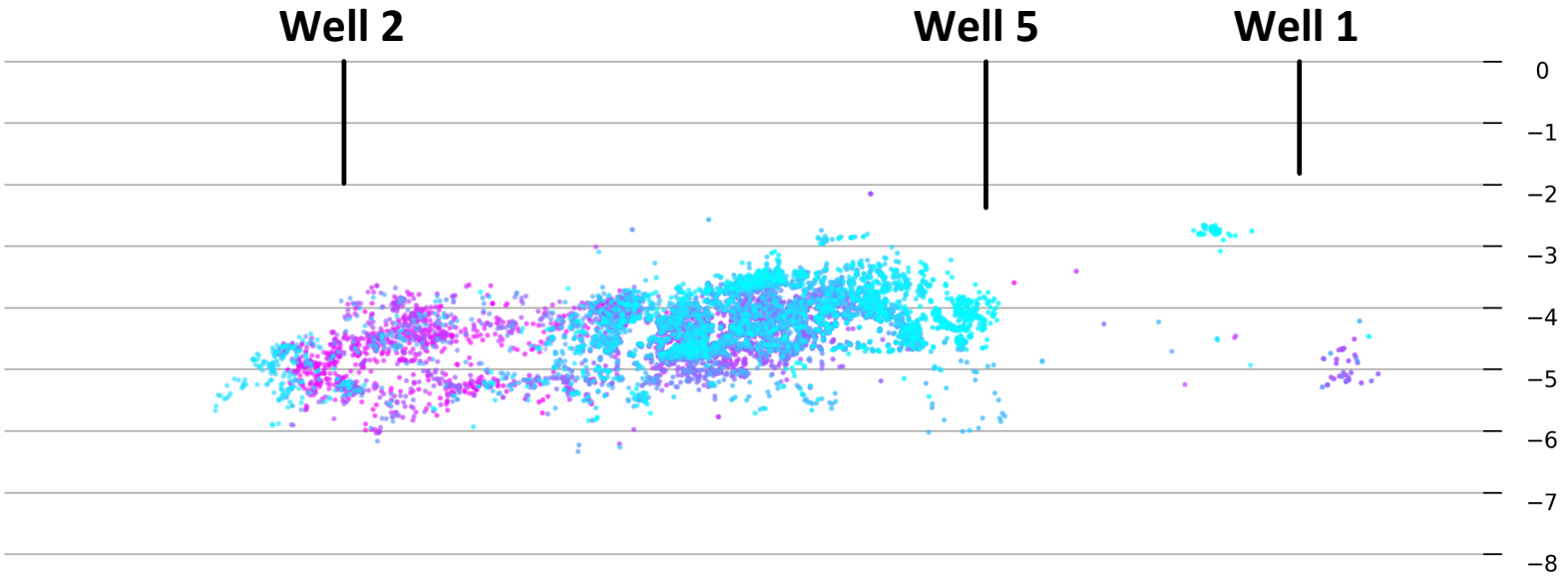

Latitude (degrees)
Depth (km)
1
2
3
4
5
6
7
8
(a)
Well 1
Enders fault
← Well 1 starts
← Well 5 starts
35.35
35.30
35.25
35.20
07/2010
08/2010
09/2010
10/2010
11/2010
12/2010
01/2011
(b)
Well 1
Enders fault
Well 1 & 5 stops →
02/2011
03/2011
(c)
Well 1
Enders fault
Well 2 stops →
04/2011
05/2011
06/2011
07/2011
08/2011
09/2011
10/2011
Park et al. (2019)



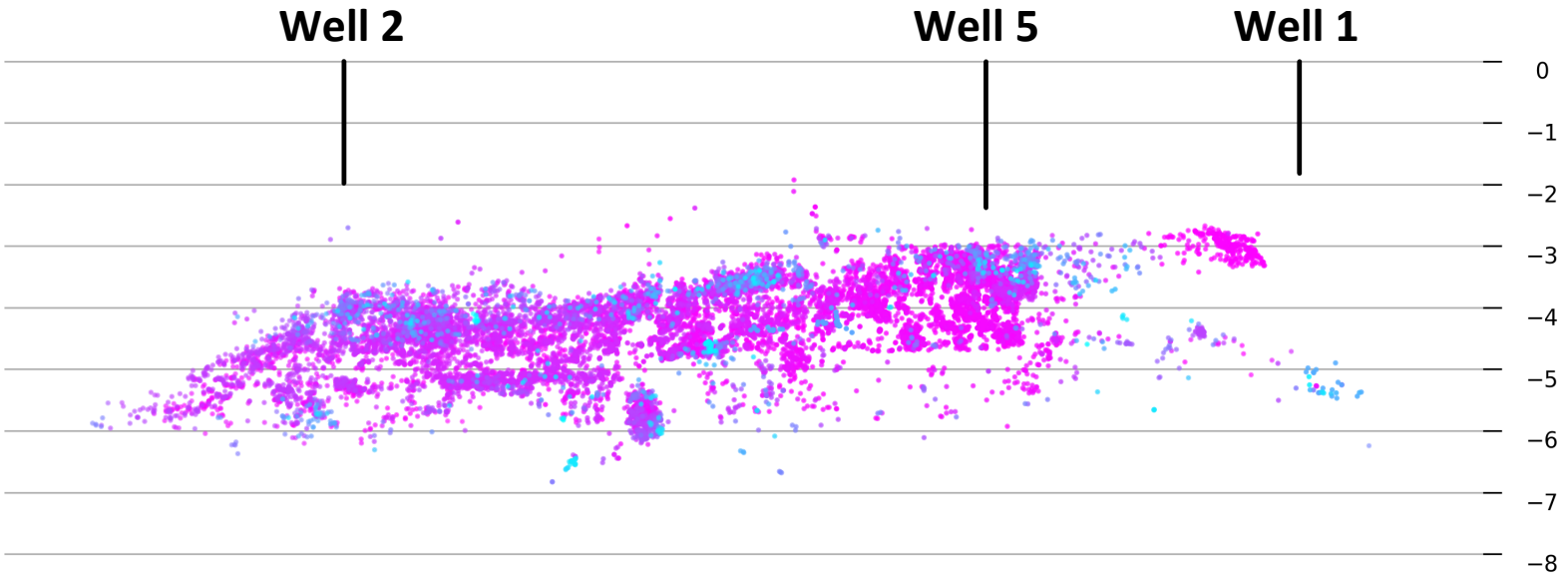
Park et al. (2019)



Park et al. (2019)

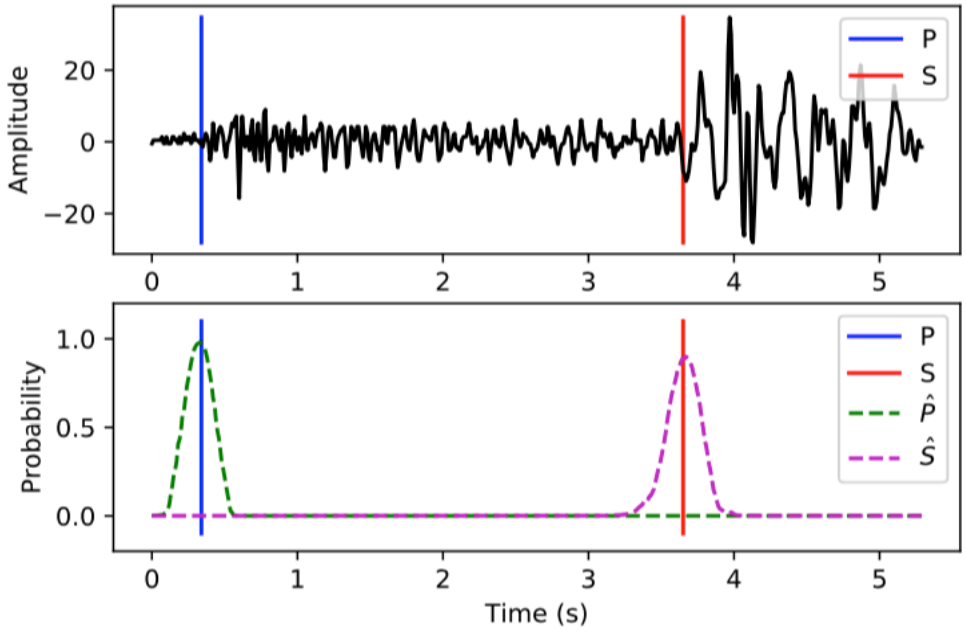


Park et al. (2019)



Park et al. (2019)

How to quantify uncertainty?



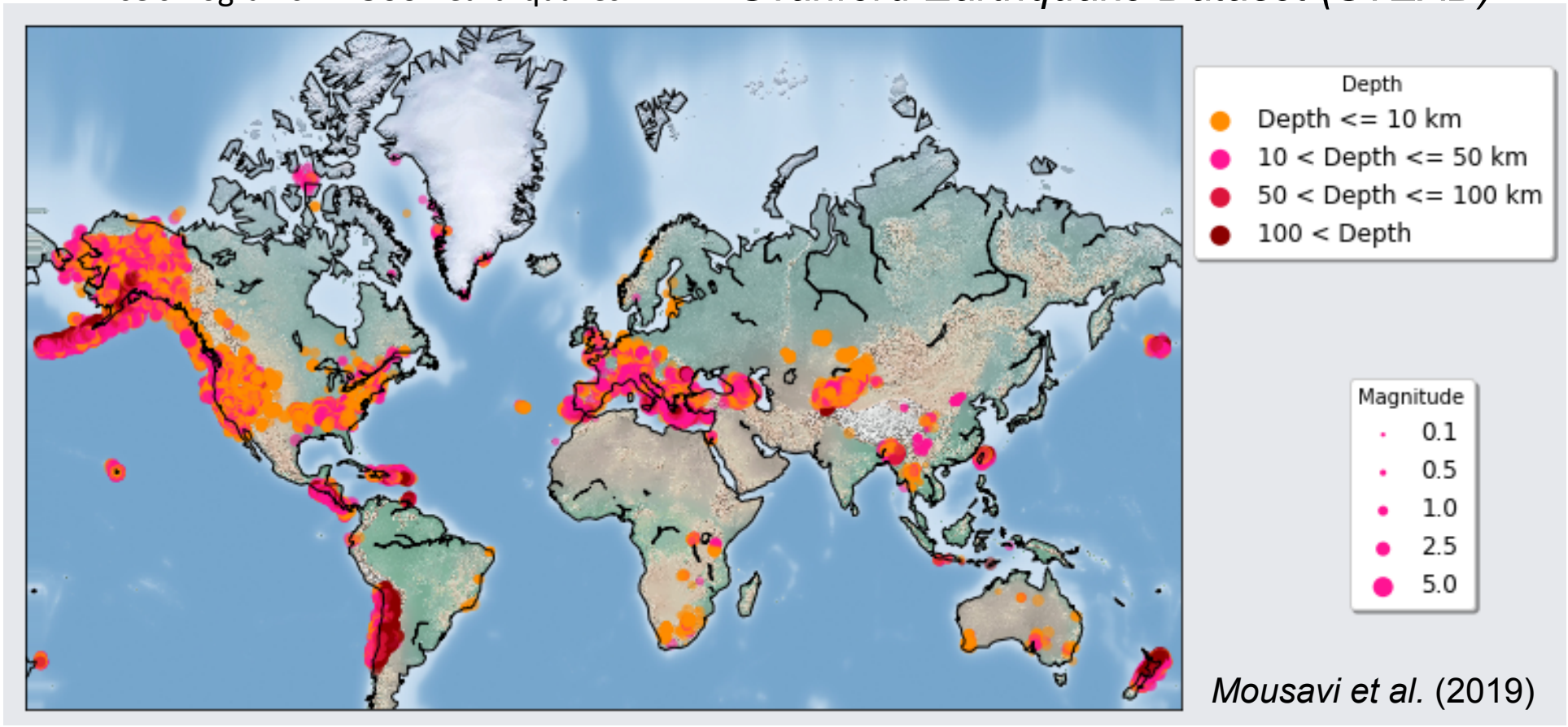
Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning

Yarin Gal YG279@CAM.AC.UK
Zoubin Ghahramani ZG201@CAM.AC.UK

Data sets and data challenges

Curated Data Sets/Benchmarks

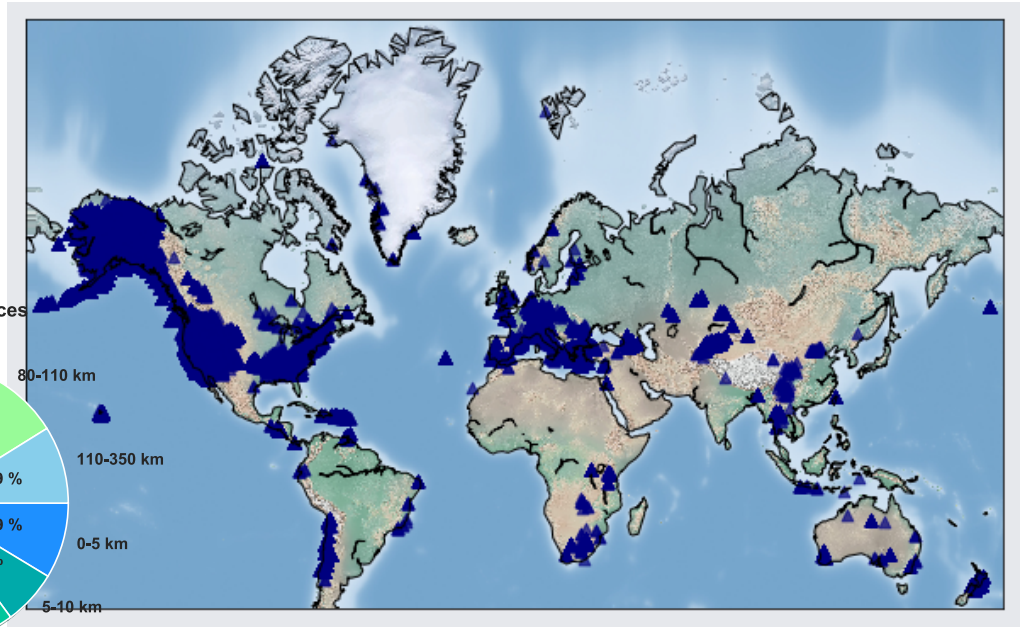
1.2 M seismograms. 500k earthquakes. *STanford Earthquake Dataset (STEAD)*

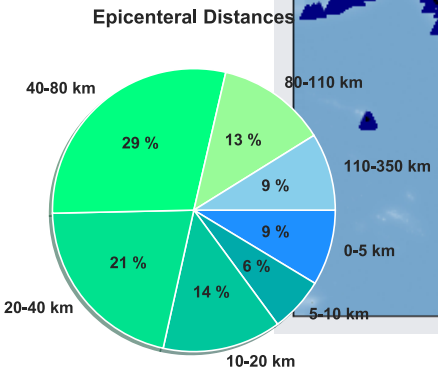


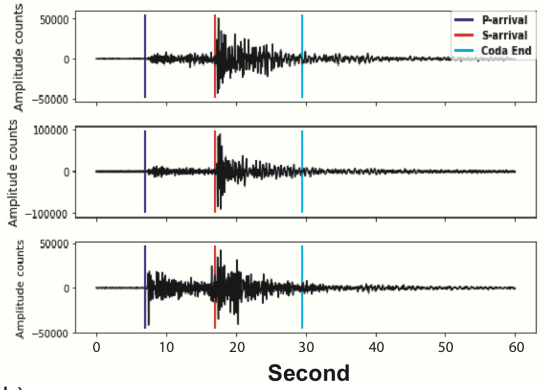
Mousavi et al. (2019)

STEAD

2,650 seismometers. **Local distances**







b)


```
{'back_azimuth_deg': 141.5,
 'category': 'earthquake',
 'coda_end_sample': 2951,
 'earthquake_depth_km': 13.9,
 'earthquake_distance_deg': 0.7026,
 'earthquake_distance_km': 77.99,
 'earthquake_id': 'ci37917624',
 'earthquake_lat': 34.001,
 'earthquake_lon': -116.9018,
 'earthquake_magnitude': 3.5,
 'earthquake_magnitude_type': 'ml',
 'earthquake_origin_time': '2017-06-25 13:53:24.760000',
 'file_name': 'ADO.CI_20170625135330_EV',
 'instrument_type': 'HH',
 'network_code': 'CI',
 'p_arrival_time': '2017-06-25 13:53:37.650',
 'p_status': 'manual',
 'p_travel_sec': 12.89,
 'p_weight': 0.46,
 's_arrival_time': '2017-06-25 13:53:47.650',
 's_status': 'manual',
 's_weight': 0.5125,
 'snr_db': [39.3, 39.5, 36.5],
 'station_elevation_m': 908.0,
 'station_lat': 34.55046,
 'station_lon': -117.43391,
 'station_name': 'ADO',
 'trace_start_time': '2017-06-25 13:53:30.650000'}
```


Signals and noise Extensive QC Additional Labels

Mousavi et al. (2019)

STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI

S. MOSTAFA MOUSAVI[1], YIXIAO SHENG[1], WEIQIANG ZHU[1], and GREGORY C. BEROZA[1]
[1]Geophysics Department, Stanford University, 397 Panama Mall, Stanford, 94305-2215, CA, United States (e-mail: mmousavi@stanford.edu)

Published in *IEEE Access*

Data-science-friendly data format

Seismology 101

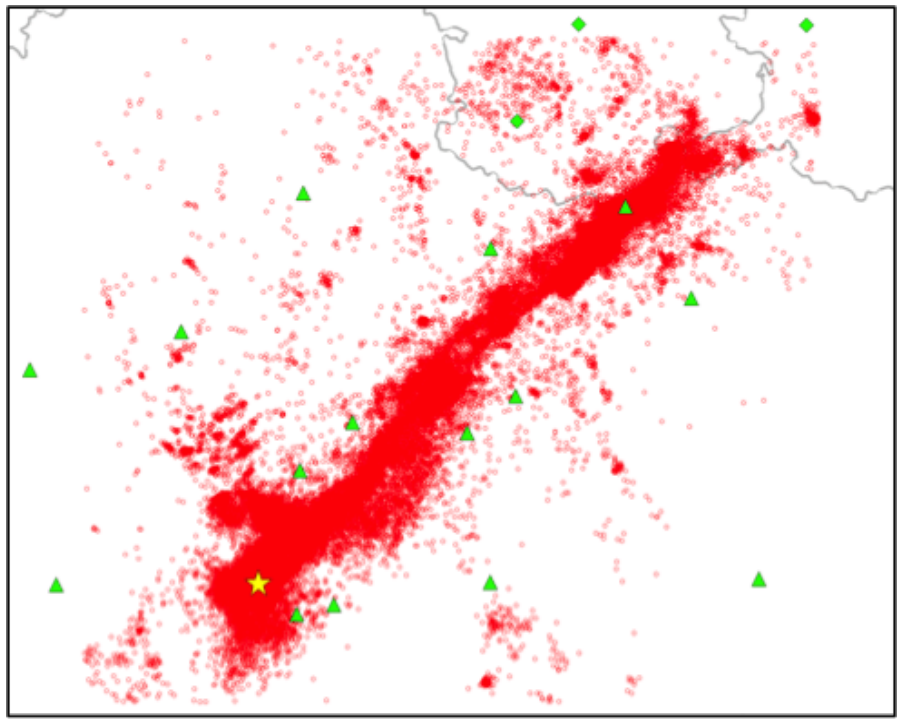
Get data scientists interested in our problems

SeismOlympics: Phase picking for 2008 Wenchuan aftershocks

1100+ teams (4000 participants)

$50,000 in prize money

Ground truth based on CEA analysts



Fang et al. (2017)

Data Description

The goal of this competition is to use seismic signals to predict the timing of laboratory earthquakes. The data comes from a well-known experimental set-up used to study earthquake physics. The `acoustic_data` input signal is used to predict the time remaining before the next laboratory earthquake (`time_to_failure`).

The training data is a single, continuous segment of experimental data. The test data consists of a folder containing many small segments. The data *within* each test file is continuous, but the test files do not represent a continuous segment of the experiment; thus, the predictions cannot be assumed to follow the same regular pattern seen in the training file.

For each `seg_id` in the test folder, you should predict a *single* `time_to_failure` corresponding to the time between the *last row of the segment* and the next laboratory earthquake.

Los Alamos NATIONAL LABORATORY EST. 1943	**Geophysics Group:** The competition builds on initial work from Bertrand Rouet-Leduc, Claudia Hulbert, and Paul Johnson. B. Rouet-Leduc prepared the data for the competition.
PennState	**Department of Geosciences:** Data are from experiments performed by Chas Bolton, Jacques Riviere, Paul Johnson and Prof. Chris Marone.
PURDUE UNIVERSITY	**Department of Physics & Astronomy:** This competition stemmed from the DOE Council workshop "Information is in the Noise: Signatures of Evolving Fracture and Fracture Networks" held March 2018 that was organized by Prof. Laura J. Pyrak-Nolte.
Department of Energy	**Office of Science, Basic Energy Sciences, Chemical Sciences, Geosciences and Biosciences Division:** The Geosciences core research.

Psi
1st place

1st place solution

posted in LANL Earthquake Prediction 5 months ago

181

Thanks a lot to the hosts of this competition and congratz to all participants and of course to my amazing teammates.

What made this competition tricky was to find a proper CV setup that you believe in as the public LB gave bad feedback for private LB. This was my first competition where this was the case and it took me a while to completely ignore public LB, but it was necessary.

I will now try to summarize some of the main points that helped us to win this competition. I am posting these elaborations in the we-form as we are a team and everyone contributed ideas and knowledge. Special thanks to @ilu000 @dott1718 @returnofsputnik @dkaraflos @pukkinming who worked hard the last few weeks on the comp.

Acoustic signal manipulation and features

As has been discussed in the forums and shown by adversarial validation, the signal had a certain time-trend that caused some issues specifically on mean and quantile based features. To partly overcome this, we added a constant noise to each 150k segment (both in train and test) by calculating `np.random.normal(0, 0.5, 150_000)`. Additionally, after noise addition, we subtracted the median of the segment.

Our features are then calculated on this manipulated signal. We mostly focused on similar features as most participants in this competition, namely finding peaks and volatility of the signal. One of our best final LGB model only used four features: (i) number of peaks of at least support 2 on the denoised signal, (ii) 20% percentile on std of rolling window of size 50, (iii) 4th and (iv) 18th Mel-frequency cepstral coefficients mean. We sometimes used a few more features (like for the NN, see below) but they are usually very similar. Those 4 are decently uncorrelated between themselves, and add good diversity. For each feature we always only considered it if it has a p-value >0.05 on a KS statistic of train vs test.

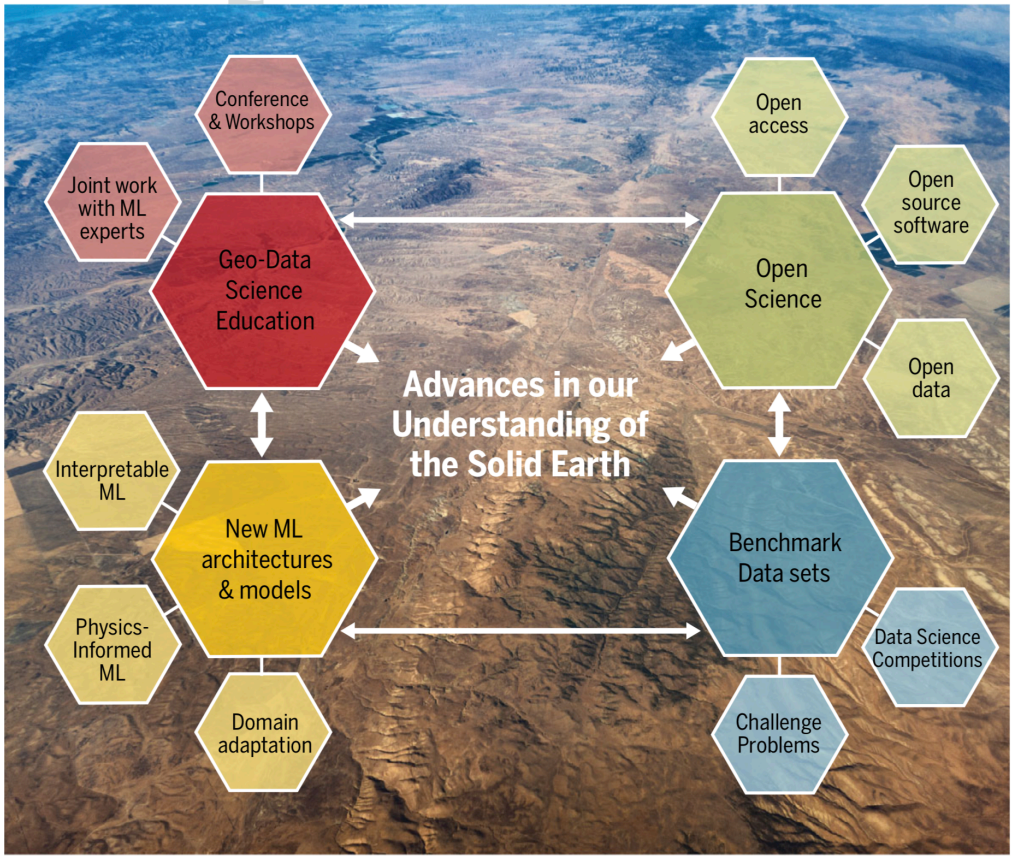
Recommendations

Accelerate progress and expand applications across geosciences

Keep up with state-of-the-art and recruit data scientists to work on our problems

Promote best practices and understand limitations

Use domain knowledge in problem solving.



Bergen et al. (2019)